DOCUMENT RESUME

ED 347 186 TM 018 644

AUTHOR Kaplan, Bruce A.; Johnson, Eugene G.
TITLE Reliability of Professionally Scored Data: NAEP-Related Issues.
PUB DATE Apr 92
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; Elementary Secondary Education; *Estimation (Mathematics); Evaluators; *Interrater Reliability; Mathematical Models; Measurement Techniques; Multiple Choice Tests; *National Surveys; Scores; *Scoring; *Standardized Tests; Test Construction; Testing Problems; *Test Reliability; Writing Tests
IDENTIFIERS Intraclass Correlation; Kappa Coefficient; *National Assessment of Educational Progress; Subjective Evaluation

ABSTRACT

Across the field of educational assessment the case has been made for alternatives to the multiple-choice item type. Most of the alternative types of items require a subjective evaluation by a rater. The reliability of this subjective rating is a key component of these types of alternative items. In this paper, measures of reliability are examined, including measures of the reliability between raters within a year, the reliability of groups of raters between years, and other parameters. Measures examined are: (1) percent exact agreement; (2) Cohen's kappa statistic; (3) intraclass correlation coefficient; and (4) minimum variance unbiased estimator of intraclass correlation. These issues have been important in the National Assessment of Educational Progress (NAEP), and are addressed in this paper through the use of NAEP writing data from 1984 through 1990. Eight data tables and 16 references are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED347186

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRUCE KAPLAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Reliability of Professionally Scored Data:

NAEP-related Issues

Bruce A. Kaplan and Eugene G. Johnson

Educational Testing Service, Princeton, New Jersey

April 24, 1992

Presented during the symposium entitled
Current Psychometric Issues in NAEP
at the American Educational Research Association Meetings
in San Francisco

TM018644

Abstract

Across the field of educational assessment the case has been made for alternatives to the multiple-choice item type. Most of the alternative types of items require a subjective evaluation by a rater. The reliability of this subjective rating is a key component of these types of alternative items. In this paper we examine measures of reliability, the reliability between raters within a year, and the reliability of groups of raters between years and other issues. These issues have been important to the National Assessment of Educational Progress (NAEP), and will be addressed in this paper through the use of NAEP writing data from 1984 through 1990.

Reliability of Professionally Scored Data: NAEP-related Issues

Bruce Kaplan and Eugene Johnson

I. Introduction

Constructed response items, open-ended items, and figural-response items are all terms that indicate change in the assessment of our students. Much of the current thought is that these types of items would be a more authentic measure than multiple choice items of what our students know. But these types of items add many challenges to the measurement field. This paper will focus on one of these challenges, the reliability of professionally scored data. We will be specifically concerned with the reliability of professionally scored data from the National Assessment of Educational Progress (NAEP).

Each of the new item types require a rater to evaluate the student's response, and make a somewhat subjective decision on the value of this response. In addition to the criteria actually being evaluated, such inappropriate factors as handwriting, neatness, and spelling may influence the raters' evaluation.

Ideally, if two raters (or for that matter the same rater at different times) are given the same paper to rate, the raters would always rate the paper the same. However, since raters are people and people are only human, exact agreement will generally not happen 100% of the time. Measures of how well raters agree include percent exact agreement and an intraclass reliability coefficient between raters. The reliability of the ratings within a particular assessment is called within-year reliability.

One of the main purposes of NAEP is to measure educational progress over time. This trend analysis requires papers scored in previous years to be compared with papers scored in the current year. If a rater in the current year is given a paper from the previous year, ideally the rater will score the paper the same way as the previous rater did. We can compute a percent exact agreement between old ratings and new ratings on the same papers, as well as a corresponding intraclass reliability coefficient. The reliability of the ratings between different assessment years is called between-year reliability.

In both within-year reliability and between-year reliability, our goals are the same; that is, to disentangle rater effects from the actual performance measure, or at least ascertain the magnitude of the rater effects. The remainder of this paper will talk about the scoring process in NAEP, provide more detail for some of the measures of reliability, and then list some ideas about how to increase reliability. The focus will be on the subject area of writing.

II. Description of Scoring Procedures in NAEP

The number of responses to open-ended items in a typical National Assessment is staggering. Each response must be scored at least once by a rater, and in some cases, as in certain writing items, an item response may be scored 6 times by different raters for 3 different types of scores. Across all subject areas assessed in 1990, there were approximately 3 million responses to open-ended problems, while in 1992 there will be over 7 million (Mullis, 1991). Table 1 displays the counts by subject area of the estimated numbers of open-ended responses to be scored in 1992.

The training of the approximately 300 raters is a major task. It is accomplished in the following manner. First, scoring guides are carefully prepared by NAEP staff in conjunction with subject area specialists. Next, ETS subject area specialists and head raters review the scoring guides and discuss examples. The raters are then trained on sample papers. This training is time-consuming and expensive, but essential to obtaining reliable ratings of open-ended responses.

The NAEP 1990 writing trend assessment consisted of six 15-minute writing prompts for each of three grades, 4, 8, and 11. These were identical to prompts given in 1984 and 1988. A student received a NAEP booklet, which consisted of a common background question block and three cognitive blocks. A booklet had either one or two writing blocks, and the remainder of the blocks contained reading items. Within a writing block there were several writing background questions and either one or two writing prompts.

All writing prompts were scored using the "primary trait" method, which focuses on the writer's effectiveness in accomplishing the specific purpose of a given task. The scoring rubric was defined such that there are five levels of task accomplishment: not rated, unsatisfactory, minimal, adequate, and elaborated. The scoring guides described these levels in detail and were specific for each writing item. The identical scoring guides were used for each assessment.

Two prompts per age group were also scored holistically. Holistic scoring evaluates the overall fluency of the respondents' writing. In addition, a subsample of one of the two prompts at each age level was scored for mechanics of writing. This scoring was intended to measure spelling, punctuation, and grammar ability. Different raters scored these papers on these three types of scoring. This paper will concentrate on the primary trait scoring, which is used to construct the writing proficiency scores. See *The NAEP 1990 Technical Report* (Johnson & Allen, 1992) and *Focusing the New Design: The NAEP 1988 Technical Report* (Johnson & Zwick, 1990) for more details.

In the 1990 assessment, for primary trait scoring the chief readers reviewed the scoring guide with ETS and NAEP subject area specialists and chose example papers for each level. The scoring guides and the examples were explained to the chief readers. Then the actual readers reviewed the sample scored responses, and conducted extensive practice scoring. For

writing trend analyses, 1988 papers were used for the training, because 1988 is NAEP's reference year for writing trend. Before actual scoring began, about 25% of the 1988 papers were mixed in with the 1990 papers and were rated concurrently. These 1988 papers, rescored by 1990 raters, give us an estimate of between-year rater reliability. Additionally, to allow the examination of within-year reliability, 25% of the 1990 papers were scored by two raters. This scoring was done blindly; that is, the second rater did not know the score given by the first rater. Ideally all the 1984 papers, 1988 papers, and the 1990 papers would have been rescored, but the cost would have been prohibitive.

III. Methods of Measurement of Rater Reliabilities

Several methods can be used to measure rater reliability, regardless of whether it is within-year or between-year rater reliability. Before discussing these, we provide a few assumptions made in NAEP writing analyses. The primary trait ratings are assigned the following numerical codes: 0=not rated, 1=unsatisfactory, 2=minimal, 3=adequate, and 4=elaborated. For writing long-term trend a booklet consists of a common background question block and three cognitive blocks. Each cognitive block is separately timed, providing 15 minutes for completion. A booklet has one or two writing blocks, and the remainder of the blocks contain reading items. Within a writing block there are several writing background questions and either one or two writing prompts. As in all other subject areas assessed by NAEP, we distinguish between two types of student nonresponse to writing tasks. The first type of nonresponse, called omit, occurs when the student chooses not to respond to the task; the second type, called not-reached, occurs when the student has not had an opportunity to respond to the task because of a lack of time. Usually, an omit is scored as wrong or partially correct, while a not-reached item is excluded from the analysis for that student. If the only prompt in the block is left blank, it is treated as an omit, and therefore given a score of zero. This is because it is believed that the student had reached the item, but chose not to respond, since there are no other cognitive items in the block. In blocks where there are two items per block, if the second item is left blank, it is treated as not-reached, with the reasoning that the student spent all his or her time on the first item. In this case, the student is excluded from the analysis of the not-reached item. If the first item is left blank, it is scored as an omit, regardless of whether the second item was answered.

Most NAEP analyses use student-level sampling weights to account for the differential selection of students into the NAEP sample. This is important when the analyses deal with student-level estimates. However, when we evaluate rater reliabilities, the analysis of interest is comparing raters, not students, so the sampling weights are not used.

Our first measure of rater reliability is percent exact agreement. Algebraically, the percent of exact agreement P_j is computed for each item j as follows:

$$P_j = 100 \times \frac{\sum_{i=1}^{N} X_i}{N}$$

where $X_i = 0$ if the raters disagree on the ith paper,
$= 1$ if the raters agree on the ith paper.

and N is the number of papers scored by two raters.

The higher the percent, the more often the raters agree. The percent of exact agreement does not take into account the possibility that two raters might assign the same rating to a paper purely by chance. Cohen (1960) introduced a statistic, called kappa, which measures the extent of agreement beyond that expected by chance. For an item with C categories let π_{hk} be the proportion of papers rated in category h by the first rater and in category k by the second. Then, the proportion of responses in which the raters agreed is $p_o = \sum_h \pi_{hh}$. If the ratings of each paper were statistically independent, the proportion of responses for which agreement is expected purely by chance is $p_c = \sum_h \pi_{h\cdot}\pi_{\cdot h}$, where $\pi_{h\cdot} = \sum_{k=1}^{C} \pi_{hk}$ is the marginal proportion of times rater 1 assigned papers to category h and π_{+h} is analogously defined for rater 2. Cohen's kappa, κ, is defined as

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

and gives the excess of observer agreement over that expected by chance as a proportion of the maximum possible disagreement expected by chance.

Both the percent of exact agreement and Cohen's kappa provide measures of the extent of rater agreement. But neither measure takes into account the amount of difference in the ratings when the raters disagree. In other words, if a paper is rated on a four-point scale, ratings of 1 and 2 show more similar judgments than ratings of 1 and 4. The percent of exact agreement treats these disagreements equally. A correlation, more precisely an intraclass correlation, takes the magnitude of the differences into account. (There is also a variant of kappa designed for ordinal categories which accounts for the magnitude of differences--see Spitzer, Cohen, Fleiss, & Endicott, 1967; Fleiss & Cohen, 1973).

For the intraclass correlation coefficient the raters are assigned randomly as the first rater and the second rater. Thus it is as likely that rater 1 be the first rater and rater 2 be the second rater, as the reverse, rater 2 being the first rater and rater 1 being the second rater. In fact, a Pearson correlation computed by including each observation twice, but reversing the rater order, provides an estimate of the intraclass correlation coefficient.

Another estimate, r(ICC), of the intraclass correlation coefficient is as follows (see Zwick, 1985):

$$r(ICC) = \frac{MSS-MSR}{MSS+MSR}$$

where,

$$MSS = \frac{1}{n-1}\left[\sum_{i=1}^{n}\frac{(\sum_{j=1}^{2}X_{ij})^2}{2} - \frac{(\sum_{i=1}^{n}\sum_{j=1}^{2}X_{ij})^2}{2\times n}\right]$$

is the between *student mean square* and,

$$MSR = \frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{2}X_{ij}^2 - \sum_{i=1}^{n}\frac{(\sum_{j=1}^{2}X_{ij})^2}{2}\right]$$

is the between *rater within student mean square*.
and X_{ij} is the ith student's score given by the jth rater, $i=1,\ldots n$ students and $j=1,2$ raters.

The estimate, r(ICC), above is commonly used by NAEP, but has a negative bias.

Olkin and Pratt (1958) proposed a minimum variance unbiased estimator of the intraclass correlation, ru(ICC). This estimator will be computed for the 1992 NAEP results, using the algorithm of Donoghue and Collins (1990).

IV. Some Examples from NAEP Writing Assessments, 1984-1990.

Tables 2 through 7 contain the rater reliabilities for several assessments. The reliabilities are calculated from results of the 1984, 1988, and 1990 writing trend assessments. Tables 2 through 6 provide item-level information by assessment year and type of reliabilities (within-year or between-year). Table 7 provides summarization over items. All tables display the following statistics for each of the three age/grade cohorts: percentage exact agreement (labeled % agree), the estimate of the intraclass correlation (r(ICC)), the estimate of the unbiased estimate of the intraclass correlation (ru(ICC)), and Cohen's kappa (Kappa). The minimum, maximum, average, and standard deviations are computed for each column. In the tables, the symbol '-' means an item was not presented to that age/grade class.

Tables 2, 4, and 6 give the within-year statistics for the 1990, 1988, and 1984 writing trend assessments, respectively. Table 3 gives the between-year statistics for the 1988 and 1990 assessments. Table 5 gives the between-year statistics for the 1984 and 1988 assessments. Table 7 contains information combined over the within-year and between-year tables.

The first observation is that there is hardly any differences between the estimates r(ICC) and ru(ICC). They differ, at most, by .002, the rank ordering of the means for both statistics by age class correspond exactly, and the correlation between the two estimates is extremely high (essentially 1). For this reason, we shall restrict further comment to r(ICC) only. While not so highly related as r(ICC) and ru(ICC), the statistics r(ICC), percent exact agreement, and kappa are all substantially related, as the correlations in Table 8 indicate. The within-age/grade correlations between percent exact agreement and kappa range from .68 to .99 with a median of .92; the correlation between kappa and r(ICC) range from .25 (with next lowest .78) to .99, also with a median of .92; and the correlations between percent exact agreement and r(ICC) range from .16 (with the next lowest .31 and .53) to .95 with a median of .78. Consequently, all four statistics (to varying degrees) impart similar information.

A pattern that is apparent from the tables is a consistent tendency for the reliability statistics of the age 9/grade 4 papers to be the largest and the reliability statistics of the age 13/grade 8 papers to be the smallest. The statistics for the ratings for the age 17/grade 11 students are intermediate, but tend to be closer to the age 9/grade 4 values. This is true for all tables except the 1988-1990 between-year tables. There might be several reasons for the 9 year-olds to have the largest reliability. There are few 9 year-old students who receive the highest score, so the range is more restricted. The 13 year-olds probably have a wider range of abilities so that the opportunity of rater disagreement is increased. The restriction of range for the age 17/grade 11 students, where there are relatively fewer students receiving the lowest score, may explain why the statistics for this age class are closer to those for the youngest age class. There also might be some effect due to the order of scoring. The 13 year-old assessment is in the fall, and their papers are scored first, followed by the 9 year-olds in the winter and then the 17 year-olds in the spring. This order of scoring could conceivably have some impact on the results.

The average overall percent exact agreement, as shown in Table 7, is 83.7 across all grades, years, and combining within and between. About 84 percent of the time, two raters agree. The average overall rater reliability is .825. This compares quite favorably with other studies where rater reliabilities are computed (see Dunbar, Koretz, & Hoover, 1991).

The average overall percent exact agreement for the within-year raters is 89.1, 80.1, 86.3, and 85.4 respectively for age 9/grade 4, age 13/grade 8, age 17/grade 11, and the average over all ages as displayed in Table 7. For the between-year raters, the average of the overall percent exact agreement is 85.6, 79.3, 78.6, and 81.1 respectively for age 9/grade 4, age 13/grade 8, age 17/grade 11, and the average over all ages. In every case the within-year percent exact agreement is greater than the respective between-year average. This is not surprising, since the between-year estimates have an added between-year component of variation.

V. Discussion

Essay scoring reliability is of considerable concern. Although the NAEP rater reliability statistics compare quite favorably with reliabilities from similar studies, we are always looking at ways to improve our methods.

A typical procedure for improving the reliabilities of student ratings is to employ multiple ratings of each paper and average the result. While this will indeed improve the reliability of the student ratings, it is a very expensive procedure, particularly in the case of NAEP, where the aim is not to obtain estimates of individual performance, but rather to obtain cost-effective estimates of the distribution of performance for subgroups of the population (such as groups defined by gender or race/ethnicity). Modelling the components of variance of an individual score based on m ratings by $\sigma_s^2 + \frac{1}{m}\sigma_r^2$, where σ_s^2 is the variance of the true student rating and σ_r^2 is the variance attributable to rater effects, the ratio of the standard error of a subgroup mean based on m ratings for each paper to the equivalent standard error based on single ratings is

$$[(\sigma_s^2 + \frac{1}{m}\sigma_r^2)/(\sigma_s^2 + \sigma_r^2)]^{1/2} = [r + \frac{1}{m}(1-r)]^{1/2} .$$

In the above, $r = \sigma_s^2/(\sigma_s^2+\sigma_r^2)$ is the reliability of a single rating.

With the reliabilities obtained by NAEP, multiple ratings are not a cost-effective way to reduce the standard errors of key statistics such as subgroup means. For example, the proportional reduction in the standard error of a subgroup mean based on double ratings is .955 for a reliability of .825, the overall average of the rater reliabilities reported in this paper. Thus, doubling the rating effort leads to only a 4.5 percent decrease in the size of the standard errors.

The between-year component of variance in the between-year rater reliabilities could be eliminated by rescoring all papers from previous years along with the current year's papers. Unfortunately, this strategy becomes enormously expensive as the number of years to be compared increases. For example, employing this strategy for the measurement of trends across the 1984, 1988, 1990, and 1992 assessments would require the scoring in 1992 of all the papers from all four assessments.

An alternative is to employ a statistical approach to adjust ratings to remove between-rater variance attributable to systematic sources of variation. For example, adjusting downward the ratings of a rater who consistently assigns higher ratings than other raters would improve the within-year reliability; adjusting the mean score of all raters at time 2 by the difference between the original mean scores at time 1 on papers rescored at time 2 and the mean of the rescored ratings of those papers at time 2 would improve the between-year reliability. The basis for these adjustments would be based on papers rescored for measuring reliability. Braun (1986,1988) discusses this approach and provides examples where

these types of adjustments lead to substantial increases in reliability. Unfortunately, experiments with linear adjustment procedures on the NAEP writing data have led to small increases in reliability, possibly at least in part because the NAEP writing reliabilities are substantially higher than the reliabilities for the studies considered by Braun.

An alternative approach for statistical adjustment of scores is a probabilistic mapping which allows for more complicated relationships than simple linear shifts and/or changes in the standard deviation. Using this approach for measurement of trend between time 1 and time 2 involves generating a predictive conditional distribution of time 2 scores for each possible score assigned at time 1. The predictive distribution of time 2 scores conditional on a time 1 score of j is generated by the observed distribution of the time 2 rescores of the papers assigned a time 1 rating of j. This is coupled with a weak prior which assigns uniform probability to each of the C possible categories available to the raters. The predictive conditional distribution is a weighted average of the observed and the prior distributions, with weight N_j (the number of papers rated j at time 1) assigned to the observed distribution and weight 1 assigned to the prior distribution. Each draw from this distribution generates a predictive value of the time 2 ratings for a time 1 rating of j. Each such predictive value can be used in place of the time 1 ratings, resulting in a probabilistic equating of the time 1 ratings to the time 2 ratings. Performing multiple draws for each paper enables the capturing of the variance due to the imprecision of the mapping, using the multiple imputation methodology of Rubin (1987). Experiments with this procedure with the 1984 and 1988 data, using the 1988 rescore of the 1984 data, showed that the group level means of the 1984 results mapped to the 1988 ratings had standard errors (including the error due to imprecision of the mapping) roughly equivalent to the standard errors of the means of the 1988 rescore of the 1984 papers. Since the precision of the remapped data was about the same as the precision of the rescored data, we opted for simplicity of reporting to use the rescored data.

Additional statistical procedures are available if summarizations of results across tasks are to be used. For example, in 1992 we will be using polytomous item response theory models (Muraki, 1990, in press) to summarize writing performance. Muraki is currently experimenting with extending these models (along the lines of the item drift models of Bock, Muraki, and Pfeiffenberger, 1988) to allow for the adjustment for systematic variation of ratings.

VI. Conclusion

The increased importance placed on professionally scored items by the educational measurement community leads directly to the increased importance of measuring and improving both within-year and between-year rater reliability. In this paper we discussed four methods of measuring rater reliability. When applied to NAEP writing trend data from 1984 to 1990, all four measures correlate highly. NAEP 1984-1990 writing trend data show within-year rater reliabilities greater than between-year rater

reliabilities, though both are quite high. In the future NAEP will explore the use of adjustments in polytomous item response theory models to better take into account rater variability.

References

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25, 275-285.

Braun, H. I. (1986, June). *Calibration of essay readers: Final report* (Program Statistics Research Technical Report No. 86-68). Princeton, NJ:, Educational Testing Service.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13(1)*, 1-18.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Donoghue, J. R., & Collins, L.M. (1990). A note on the unbiased estimation of the intraclass correlation. *Psychometrika, 55(1)*, 159-164.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991, April). Quality control in the development and use of performance assessments. In *Technical and conceptual issues in the evaluation of performance assessments*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Chicago.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613-619.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report*.(No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Allen, N. A. (1992). *The NAEP 1990 technical report* (No. 21-TR-20). Princeton, NJ: Educational Testing Service; Washington, DC: National Center for Education Statistics.

Mullis, I. V. S. (1991). *The NAEP guide: A description of the content and methods of the 1990 and 1992 assessments* (No. 21-TR-01). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59-71.

Muraki, E. (in press). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*.

Olkin, I. & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29*, 202-211.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Spitzer, R. L., Cohen, J., Fleiss, J. L., and Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Arch. General Psychiatry*, 18, 83-87.

Zwick, R. (1985). *Reliability coefficients for writing data*. Technical memorandum.

Table 1
1992 NAEP Estimates of Sample Size

Cross-sectional				
Sample Type	Age/Grade	Subject	Students	Open-ended Responses
National	9/ 4	Math	12,500	199,000
		Reading	10,000	121,000
		Writing	11,000	25,000
State	-/ 4	Math	120,000	2,000,000
		Reading	120,000	1,500,000
National	13/ 8	Math	12,500	192,500
		Reading	15,000	263,000
		Writing	15,000	37,000
State	-/ 8	Math	120,000	2,000,000
National	17/12	Math	12,500	211,000
		Reading	18,000	260,000
		Writing	18,000	42,000
Long-Term Trend				
Sample Type	Age/Grade	Subject	Students	Open-ended Responses
National	9/ 4	R/W	5,200	38,000
		M/S	6,000	62,000
	13/ 8	R/W	5,200	49,000
		M/S	6,000	60,000
	17/11	R/W	5,200	51,000
		M/S	4,000	64,000

Table 2
1990 Within-Year Rater Reliabilities

		Age 9/Grade 4				Age 13/Grade 8				Age 17/Grade 11			
	NAEP ID	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa
1	N000302	-	-	-	-	76.7	.725	.726	.547	76.3	.780	.780	.635
2	N000402	-	-	-	-	72.1	.669	.670	.516	76.7	.726	.727	.562
3	N000502	-	-	-	-	66.1	.560	.561	.385	-	-	-	-
4	N000602	89.6	.843	.844	.791	86.8	.761	.762	.715	-	-	-	-
5	N000902	92.1	.925	.925	.868	80.7	.833	.834	.706	-	-	-	-
6	N001002	78.5	.719	.720	.602	75.9	.716	.716	.586	81.6	.816	.816	.692
7	N007602	78.2	.773	.774	.632	-	-	-	-	-	-	-	-
8	N014702	82.4	.860	.861	.738	-	-	-	-	-	-	-	-
9	N014802	75.2	.822	.822	.633	-	-	-	-	-	-	-	-
10	N018002	-	-	-	-	-	-	-	-	71.8	.748	.748	.555
11	N019002	-	-	-	-	-	-	-	-	84.6	.828	.829	.707
12	N021002	-	-	-	-	-	-	-	-	75.6	.784	.785	.624
	Minimum	75.2	.719	.720	.602	66.1	.560	.561	.385	71.8	.726	.727	.555
	Maximum	92.1	.925	.925	.878	86.8	.833	.834	.715	84.6	.828	.829	.707
	Average	82.7	.824	.824	.711	76.4	.711	.712	.576	77.8	.780	.781	.629
	Standard Deviations	6.79	.071	.071	.106	7.09	.092	.092	.124	4.58	.039	.039	.063

Table 3
1988-1990 Between-Year Rater Reliabilities

		Age 9/Grade 4				Age 13/Grade 8				Age 17/Grade 11			
	NAEP ID	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa
1	N000302	-	-	-	-	83.0	.805	.806	.710	71.6	.779	.780	.576
2	N000402	-	-	-	-	83.5	.779	.780	.688	78.9	.693	.694	.533
3	N000502	-	-	-	-	80.6	.700	.701	.603	-	-	-	-
4	N000602	91.1	.897	.897	.845	92.6	.865	.865	.823	-	-	-	-
5	N000902	89.0	.900	.901	.821	82.0	.789	.790	.725	-	-	-	-
6	N001002	76.9	.776	.777	.615	75.4	.754	.755	.601	81.1	.811	.811	.688
7	N007602	80.5	.741	.742	.675	-	-	-	-	-	-	-	-
8	N014702	88.5	.889	.889	.827	-	-	-	-	-	-	-	-
9	N014802	83.7	.890	.891	.768	-	-	-	-	-	-	-	-
10	N018002	-	-	-	-	-	-	-	-	73.2	.748	.749	.542
11	N019002	-	-	-	-	-	-	-	-	85.5	.856	.857	.736
12	N021002	-	-	-	-	-	-	-	-	78.2	.764	.765	.650
	Minimum	76.9	.741	.742	.615	75.4	.700	.701	.601	71.6	.693	.694	.533
	Maximum	91.1	.900	.901	.845	92.6	.865	.865	.823	85.5	.856	.857	.736
	Average	85.0	.849	.850	.759	82.9	.782	.783	.692	78.1	.775	.776	.621
	Standard Deviations	5.53	.071	.071	.094	5.60	.055	.054	.083	5.11	.056	.055	.083

Table 4
1988 Within-Year Rater Reliabilities

		Age 9/Grade 4				Age 13/Grade 8				Age 17/Grade 11			
	NAEP ID	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa
1	N000302	-	-	-	-	85.4	.821	.822	.740	90.8	.931	.932	.861
2	N000402	-	-	-	-	79.9	.685	.686	.582	93.1	.864	.865	.808
3	N000502	-	-	-	-	76.1	.644	.645	.517	-	-	-	-
4	N000602	97.1	.985	.985	.950	93.5	.920	.920	.867	-	-	-	-
5	N000902	93.5	.950	.950	.897	87.0	.886	.887	.804	-	-	-	-
6	N001002	90.3	.916	.917	.848	75.3	.690	.691	.565	89.3	.889	.890	.815
7	N007602	87.5	.875	.877	.793	-	-	-	-	-	-	-	-
8	N014702	94.3	.947	.947	.912	-	-	-	-	-	-	-	-
9	N014802	91.8	.952	.953	.884	-	-	-	-	-	-	-	-
10	N018002	-	-	-	-	-	-	-	-	89.9	.929	.929	.844
11	N019002	-	-	-	-	-	-	-	-	92.3	.922	.923	.859
12	N021002	-	-	-	-	-	-	-	-	84.9	.867	.868	.767
	Minimum	87.5	.875	.877	.793	75.3	.644	.645	.517	84.9	.864	.865	.808
	Maximum	97.1	.985	.985	.950	93.5	.920	.920	.867	93.1	.931	.932	.929
	Average	92.4	.938	.938	.881	82.9	.774	.775	.679	90.1	.900	.901	.826
	Standard Deviations	3.34	.038	.037	.054	7.05	.117	.116	.144	2.90	.031	.031	.036

Table 5
1984-1988 Between-Year Rater Reliabilities

		Age 9/Grade 4				Age 13/Grade 8				Age 17/Grade 11			
	NAEP ID	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa
1	N000302	-	-	-	-	75.2	.727	.728	.594	74.1	.723	.723	.592
2	N000402	-	-	-	-	72.8	.584	.584	.472	78.4	.708	.709	.577
3	N000502	-	-	-	-	71.5	.574	.574	.411	-	-	-	-
4	N000602	92.1	.931	.932	.870	83.9	.754	.755	.676	-	-	-	-
5	N000902	90.4	.917	.917	.853	80.6	.835	.834	.714	-	-	-	-
6	N001002	77.4	.818	.819	.664	69.8	.686	.687	.520	84.3	.874	.874	.742
7	N007602	85.1	.807	.808	.740	-	-	-	-	-	-	-	-
8	N014702	87.4	.883	.883	.801	-	-	-	-	-	-	-	-
9	N014802	84.2	.857	.857	.769	-	-	-	-	-	-	-	-
10	N018002	-	-	-	-	-	-	-	-	77.1	.826	.826	.638
11	N019002	-	-	-	-	-	-	-	-	84.3	.848	.848	.719
12	N021002	-	-	-	-	-	-	-	-	75.5	.813	.813	.642
	Minimum	77.4	.807	.808	.664	69.8	.574	.574	.411	74.1	.708	.709	.577
	Maximum	92.1	.931	.932	.870	83.9	.834	.834	.714	84.3	.874	.874	.742
	Average	86.1	.869	.869	.783	75.6	.693	.694	.565	79.0	.799	.799	.652
	Standard Deviations	5.23	.051	.051	.076	5.52	.101	.101	.118	4.39	.068	.068	.066

Table 6
1984 Within-Year Rater Reliabilities

		Age 9/Grade 4				Age 13/Grade 8				Age 17/Grade 11			
	NAEP ID	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa	% agree	r(ICC)	ru(ICC)	Kappa
1	N000302	-	-	-	-	86.1	.865	.865	.768	90.2	.914	.914	.849
2	N000402	-	-	-	-	82.9	.774	.775	.675	92.2	.901	.901	.859
3	N000502	-	-	-	-	79.0	.727	.728	.591	-	-	-	-
4	N000602	93.0	.930	.931	.878	89.1	.861	.861	.792	-	-	-	-
5	N000902	95.6	.969	.969	.933	84.2	.878	.878	.773	-	-	-	-
6	N001002	89.9	.926	.926	.853	79.4	.836	.836	.681	90.1	.928	.928	.838
7	N007602	92.3	.905	.905	.863	-	-	-	-	-	-	-	-
8	N014702	92.5	.940	.941	.883	-	-	-	-	-	-	-	-
9	N014802	89.2	.909	.909	.845	-	-	-	-	-	-	-	-
10	N018002	-	-	-	-	-	-	-	-	91.2	.933	.934	.862
11	N019002	-	-	-	-	-	-	-	-	92.4	.932	.932	.865
12	N021002	-	-	-	-	-	-	-	-	89.0	.910	.910	.838
	Minimum	89.2	.905	.905	.845	79.0	.727	.728	.591	89.0	.901	.901	.838
	Maximum	95.6	.969	.969	.933	89.1	.878	.878	.792	92.4	.933	.934	.865
	Average	92.1	.930	.930	.876	83.5	.824	.824	.713	90.9	.920	.920	.852
	Standard Deviations	2.30	.023	.023	.031	3.90	.060	.060	.078	1.32	.013	.013	.012

Table 7
Within-Year vs. Between-Year Rater Reliabilities
Summarized Across Years

		Age 9/Grade 4			Age 13/Grade 8			Age 17/Grade 11			Across Ages		
		% agree	r(ICC)	Kappa	% agree	r(ICC)	Kappa	% agree	r(ICC)	Kappa	% agree	r(ICC)	Kappa
Within	Minimum	75.2	.719	.602	66.1	.560	.385	71.8	.726	.555	66.1	.560	.385
	Maximum	97.1	.985	.950	93.5	.920	.867	93.1	.933	.929	97.1	.985	.950
	Average	89.1	.897	.823	80.9	.770	.656	86.3	.867	.769	85.4	.845	.749
Between	Minimum	76.9	.741	.615	69.8	.574	.411	74.1	.693	.533	69.8	.574	.411
	Maximum	92.1	.931	.870	92.6	.865	.823	85.5	.874	.742	92.6	.931	.870
	Average	85.6	.859	.771	79.3	.738	.629	78.6	.787	.637	81.1	.795	.679
Overall	Minimum	75.2	.719	.602	66.1	.560	.385	71.6	.693	.533	66.1	.560	.385
	Maximum	97.1	.985	.950	93.5	.920	.867	93.1	.933	.929	97.1	.985	.950
	Average	87.7	.882	.802	80.3	.757	.645	83.2	.835	.716	83.7	.825	.721

Table 8
Correlations Among Methods of Measuring Rater Reliability

Type/year	Methods being correlated	Age 9/ Grade 4	Age 13/ Grade 8	Age 17/ Grade 11
Within-year/1990	% agree vs kappa	.952	.956	.888
	% agree vs r(ICC)	.725	.854	.793
	r(ICC) vs kappa	.900	.943	.972
Within-year/1988	% agree vs kappa	.990	.970	.676
	% agree vs r(ICC)	.805	.953	.305
	r(ICC) vs kappa	.979	.997	.779
Within-year/1984	% agree vs kappa	.943	.878	.901
	% agree vs r(ICC)	.767	.657	.159
	r(ICC) vs kappa	.918	.933	.251
Between-years/ 1988-1990	% agree vs kappa	.983	.911	.783
	% agree vs r(ICC)	.846	.782	.526
	r(ICC) vs kappa	.903	.936	.891
Between-years/ 1984-1988	% agree vs kappa	.981	.861	.844
	% agree vs r(ICC)	.847	.730	.661
	r(ICC) vs kappa	.925	.966	.927